

# A Multivariate Graphical Stochastic Volatility Model

Yuan Cheng and Alex Lenkoski\*

*Heidelberg University, Germany*

March 3, 2013

## Abstract

The Gaussian Graphical Model (GGM) is a popular tool for incorporating sparsity into joint multivariate distributions. The G-Wishart distribution, a conjugate prior for precision matrices satisfying general GGM constraints, has now been in existence for over a decade. However, due to the lack of a direct sampler, its use has been limited in hierarchical Bayesian contexts, relegating mixing over the class of GGMs mostly to situations involving standard Gaussian likelihoods. Recent work, however, has developed methods that couple model and parameter moves, first through reversible jump methods and later by direct evaluation of conditional Bayes factors and subsequent resampling. Further, methods for avoiding prior normalizing constant calculations—a serious bottleneck and source of numerical instability—have been proposed. We review and clarify these developments and then propose a new methodology for GGM comparison that blends many recent themes. Theoretical developments and computational timing experiments reveal an algorithm that has limited computational demands and dramatically improves on computing times of existing methods. We conclude by developing a parsimonious multivariate stochastic volatility model that embeds GGM uncertainty in a larger hierarchical framework. The method is shown to be capable of adapting to the extreme swings in market volatility experienced in 2008 after the collapse of Lehman Brothers, offering considerable improvement in posterior predictive distribution calibration.

---

\* *Corresponding author address:* Alex Lenkoski, Institute of Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany  
E-mail: alex.lenkoski@uni-heidelberg.de

# 1 Introduction

The Gaussian graphical model (GGM) has received widespread consideration (see Jones et al., 2005) and estimators obeying graphical constraints in standard Gaussian sampling were proposed as early as Dempster (1972). Initial incorporation of GGMs in Bayesian estimation largely focus on decomposable graphs (Dawid and Lauritzen, 1993), since prior distributions factorize into products of Wishart distributions. Roverato (2002) proposes a generalized extension of the Hyper-Inverse Wishart distribution for covariance matrices  $\Sigma$  over arbitrary graphs. Atay-Kayis and Massam (2005) turn this into a prior specified for precision matrices  $\mathbf{K}$  and outline a Monte Carlo (MC) method that enables pairwise model comparisons. Following Letac and Massam (2007) and Rajaratnam et al. (2008), Lenkoski and Dobra (2011) term this distribution the G-Wishart, and propose computational improvements to direct model comparison and model search.

A number of samplers for precision matrices under a G-Wishart distribution have been proposed. These involve either block Gibbs sampling (Piccioni, 2000), Metropolis-Hastings (MH) moves (Mitsakakis et al., 2011; Dobra and Lenkoski, 2011; Dobra et al., 2011), or rejection sampling (Wang and Carvalho, 2010). Dobra et al. (2011) shows that the rejection sampler of Wang and Carvalho (2010) suffers from extremely low acceptance probabilities in even moderate dimensions. Wang and Li (2012) conclusively show that block Gibbs sampling is both computationally more efficient and exhibits considerably less autocorrelation than the MH methods.

The block Gibbs sampler provides a Markov chain Monte Carlo (MCMC) sample. When the likelihood assumes standard Gaussian sampling, determining posterior expectations of  $\mathbf{K}$  can technically be performed as in Lenkoski and Dobra (2011), whereby model probabilities are first directly assessed via stochastic search, and model averaged samples are then collected using block Gibbs over each model. However, when the GGM is specified over latent data in a hierarchical Bayesian framework, such an approach is no longer valid. This is due to the use of the matrix  $\mathbf{K}$  in updating other hyperparameters as well as its involvement in updating the latent Gaussian factors.

Dobra and Lenkoski (2011) propose a reversible jump MCMC (which for brevity we refer to as RJ) method (Green, 1995) that simultaneously updates the GGM and its associated  $\mathbf{K}$ , and embed the GGM in a semiparametric Gaussian copula. Dobra et al. (2011) expand the RJ algorithm and show how GGMs may be used to model dependent random effects

in a generalized linear model context, focusing on lattice data. Wang and Li (2012) utilize conditional properties of G-Wishart variates that enables model moves through calculation of a conditional Bayes factor (CBF) (Dickey and Gunel, 1978) and subsequently update  $\mathbf{K}$  through direct Gibbs sampling. Wang and Li (2012) also explore the use of a double MH algorithm (Liang, 2010) to avoid the computationally expensive and numerically unstable MC approximation of normalizing constants proposed by Atay-Kayis and Massam (2005).

We investigate an alternative method for simultaneously updating the GGM and associated  $\mathbf{K}$  in hierarchical Bayesian settings. Our method is built on the framework outlined in Wang and Li (2012), but blends several of the developments above to yield an algorithm with considerably less computational cost. We show that use of a CBF on the Cholesky decomposition of a permuted version of  $\mathbf{K}$ , originally proposed in an early version of Wang and Li (2012), enables fast model moves; use of methods for sparse Cholesky decompositions (Rue, 2001) further reduce computational overhead. We then show that while both Dobra et al. (2011) and Wang and Li (2012) indicate the random walk MH sampler of Mitsakakis et al. (2011) is not suitable for posterior sampling, it is especially useful in the context of the double MH algorithm. We are therefore able to specify a model move algorithm which requires little computational effort and exhibits no numerical instability.

Simulation experiments compare our new algorithm to the algorithm of Wang and Li (2012) (which we refer to as WL). Both methods perform equally well at determining posterior distributions. However, we show that while the WL approach is theoretically appealing, it suffers significant computational overhead on account of many matrix inversions. By contrast, our new approach exhibits a dramatic improvement in computation time.

We conclude with an example of how GGMs may be embedded in hierarchical Bayesian models. GGMs have been shown to yield parsimonious joint distributions useful in financial applications (Carvalho and West, 2007; Rodriguez et al., 2011). However, existing studies have largely ignored heteroskedasticity in financial data and relied on datasets taken over periods with relatively little financial turmoil. To address this, we propose a parsimonious multivariate stochastic volatility model that incorporates GGM uncertainty. We then model stock returns for 20 assets during the period surrounding the financial crash of 2008. We show that in the periods leading up to the crash, and 9 months after the most turbulent period, this method yields no improvement over an approach that does not model heteroskedasticity. However, during the period of heightened volatility, our new model is able to adapt quickly and yields considerably more calibrated predictive distributions.

The article is organized as follows. In Section 2 we review the G-Wishart distribution, establish results necessary for CBF calculations and describe the block Gibbs sampler. Section 3 conducts a simulation study showing the computational advantage gained by our new algorithm. In Section 4 we describe our multivariate graphical stochastic volatility model and give results over the data mentioned above. We conclude in Section 5.

## 2 The G-Wishart Distribution

### 2.1 Review of Basic G-Wishart Properties

Suppose that we collect data  $\mathcal{D} = \{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(n)}\}$  such that  $\mathbf{Z}^{(j)} \sim \mathcal{N}_p(0, \mathbf{K}^{-1})$  independently for  $j \in \{1, \dots, n\}$ , where  $\mathbf{K} \in \mathbf{P}$ , the space of  $p \times p$  positive definite matrices. This sample has likelihood

$$pr(\mathcal{D}|\mathbf{K}) = (2\pi)^{-np/2} |\mathbf{K}|^{n/2} \exp\left(-\frac{1}{2}\langle \mathbf{K}, \mathbf{U} \rangle\right),$$

where  $\langle A, B \rangle = \text{tr}(A'B)$  denotes the trace inner product and  $\mathbf{U} = \sum_{i=1}^n \mathbf{Z}^{(i)} \mathbf{Z}^{(i)'}$ .

Further suppose that  $G = (V, E)$  is a GGM where  $V = \{1, \dots, p\}$  and  $E \subset V \times V$ . We will slightly abuse notation throughout, by writing  $(i, j) \in G$  to indicate that the edge  $(i, j)$  is in the edge set  $E$ . Associated with  $G$  is a subspace  $\mathbf{P}_G \subset \mathbf{P}$  such that  $\mathbf{K} \in \mathbf{P}_G$  implies that  $\mathbf{K} \in \mathbf{P}$  and  $K_{ij} = 0$  whenever  $(i, j) \notin G$ . The G-Wishart distribution (Roverato, 2002; Atay-Kayis and Massam, 2005)  $\mathcal{W}_G(\delta, \mathbf{D})$  assigns prior probability to  $\mathbf{K} \in \mathbf{P}_G$  as

$$pr(\mathbf{K}|\delta, \mathbf{D}, G) = \frac{1}{I_G(\delta, \mathbf{D})} |\mathbf{K}|^{(\delta-2)/2} \left(-\frac{1}{2}\langle \mathbf{K}, \mathbf{D} \rangle\right) \mathbf{1}_{\mathbf{K} \in \mathbf{P}_G}.$$

The normalizing constant  $I_G$  is in general not known to have an explicit form, and Atay-Kayis and Massam (2005) propose an MC approximation for this factor. Furthermore, the G-Wishart is conjugate and thus  $pr(\mathbf{K}|\mathcal{D}, G) = \mathcal{W}_G(\delta + n, \mathbf{D}^*)$  where  $\mathbf{D}^* = \mathbf{D} + \mathbf{U}$ .

Let  $\Phi$  be the upper triangular matrix such that  $\Phi'\Phi = \mathbf{K}$ , the Cholesky decomposition. Rue (2001) notes that we may associate with  $G$  another graph  $F$ , called the *fill-in* graph, such that  $G \subset F$ ,  $\Phi_{ij} = 0$  when  $(i, j) \notin F$  and

$$\Phi_{ij} = -\frac{1}{\Phi_{ii}} \sum_{l=1}^i \Phi_{li} \Phi_{lj} \quad (1)$$

when  $i < j$  and  $(i, j) \in F \setminus G$ . Rue (2001) outlines a straightforward method for constructing a graph  $F$  from  $G$  and explains how use of node reordering software can minimize fill-in.

Roverato (2002) shows that if  $K \sim \mathcal{W}_G(\delta, \mathbf{D})$  then

$$pr(\Phi|\delta, \mathbf{D}, G) = \prod_{i=1}^p \Phi_{ii}^{\delta+\nu_i^G-1} \exp\left(-\frac{1}{2}\langle \Phi' \Phi, \mathbf{D} \rangle\right), \quad (2)$$

where  $\nu_i^G$  is number of nodes in  $\{i+1, \dots, p\}$  that are connected to node  $i$  in  $G$ . We especially note that if  $\mathbf{K} \sim \mathcal{W}_G(\delta, \mathbb{I}_p)$ , then

$$pr(\Phi|\delta, \mathbb{I}_p, G) = \exp\left(-\frac{1}{2} \sum_{(i,j) \in F} \Phi_{ij}^2\right) \prod_{i=1}^p \Phi_{ii}^{\delta+\nu_i^G-1} \exp\left(-\frac{1}{2}\Phi_{ii}^2\right). \quad (3)$$

## 2.2 Sampling Methods

We review two MCMC methods for approximate sampling from a  $\mathcal{W}_G(\delta, \mathbf{D})$ . See Dobra et al. (2011) and Wang and Li (2012) for more detailed reviews of the many methods that have been proposed.

Let  $\mathcal{C}$  denote the cliques of  $G$ . In the following, we consider a clique to be a maximally complete subgraph, though Wang and Li (2012) note that this can be relaxed to any complete subgraph. Piccioni (2000) shows that for any  $C \in \mathcal{C}$ ,

$$K_C - K_{C, V \setminus C} K_{V \setminus C}^{-1} K_{V \setminus C, C} \sim \mathcal{W}(\delta, D_C), \quad (4)$$

where  $\mathcal{W}$  denotes a standard Wishart variate. The expression (4) thereby gives the full conditionals of  $\mathcal{W}_G(\delta, \mathbf{D})$ . The block Gibbs sampler thus cycles through  $\mathcal{C}$ , updating each component using (4). Wang and Li (2012) convincingly show that for posterior inference of  $\mathcal{W}_G(\delta + n, \mathbf{D}^*)$  the block Gibbs sampler outperforms all other proposed methods, both in terms of computing time and mixing. The authors also provide a useful review of the algorithm and indicate its broad flexibility. Throughout, we use the block Gibbs sampler for updating the matrix  $\mathbf{K}$  in the posterior.

Both Dobra et al. (2011) and Wang and Li (2012) show that the random walk MH (RWMH) algorithm of Mitsakakis et al. (2011) is unsuitable for posterior inference. However, we show below that it is especially effective to use in the double MH algorithm. In particular, suppose that we wish to sample from  $\mathcal{W}_G(\delta, \mathbb{I}_p)$  and  $\mathbf{K}$  is the current state of an MCMC chain. Then the RWMH algorithm performs the following

1. Determine  $\Phi$  from  $\mathbf{K}$
2. Propose  $\Psi$ :
  - a. Sample  $c \sim \chi_{\delta+\nu_i^G}^2$  and set  $\Psi_{ii} = c^{1/2}$
  - b. Sample  $\Psi_{ij} \sim \mathcal{N}(0, 1)$  for  $(i, j) \in G$
  - c. Complete  $\Psi$  using (1) for all  $(i, j) \in F \setminus G, i < j$
3. Compute

$$\alpha = \exp \left( -\frac{1}{2} \sum_{(i,j) \in F \setminus G} (\Psi_{ij}^2 - \Phi_{ij}^2) \right) \quad (5)$$

4. With probability  $\min\{\alpha, 1\}$  set  $\mathbf{K} = \Psi' \Psi$

The appeal of the RWMH algorithm in sampling from  $\mathcal{W}_G(\delta, \mathbb{I}_p)$  is the simplicity of the factor in (5). Through the use of node reordering software, which minimizes the size of  $F \setminus G$ , this expression may require few calculations. While the algorithm does not perform well in the posterior, and the calculation (5) as well as step (2.c) become more involved when  $\mathbf{D} \neq \mathbb{I}_p$  we show below that in the particular case of the double MH algorithm using the prior  $\mathcal{W}_G(\delta, \mathbb{I}_p)$ , this method is extremely useful.

## 2.3 Conditional Bayes Factors

Prior to Wang and Li (2012), model moves between two graphs  $G$  and  $G'$  focused on approximating the ratio

$$\frac{pr(G|\mathcal{D})}{pr(G'|\mathcal{D})} = \frac{pr(\mathcal{D}|G)}{pr(\mathcal{D}|G')} \times \frac{pr(G)}{pr(G')}, \quad (6)$$

first through MC (Atay-Kayis and Massam, 2005; Jones et al., 2005), then a combination of MC and Laplace approximation (Lenkoski and Dobra, 2011) and ultimately through RJ (Dobra and Lenkoski, 2011; Dobra et al., 2011).

Suppose that  $G \subset G'$  which differ only by the edge  $e = (i, j) \in G'$  and that  $\mathbf{K} \in \mathbf{P}_G$ . Let  $\mathbf{K}^{-e} = \mathbf{K} \setminus \{K_{ij}, K_{ji}, K_{jj}\}$ . In lieu of (6), Wang and Li (2012) consider ratios of the form

$$\frac{pr(G|\mathbf{K}^{-e}, \mathcal{D})}{pr(G'|\mathbf{K}^{-e}, \mathcal{D})} = \frac{pr(\mathcal{D}, \mathbf{K}^{-e}|G)}{pr(\mathcal{D}, \mathbf{K}^{-e}|G')} \times \frac{pr(G)}{pr(G')} \quad (7)$$

which are related to the conditional Bayes factors (CBFs) of Dickey and Gunel (1978).

Using properties related to the form (4) Wang and Li (2012) show that

$$\frac{pr(\mathcal{D}, \mathbf{K}^{-e}|G)}{pr(\mathcal{D}, \mathbf{K}^{-e}|G')} = H(\delta + n, e, \mathbf{K}^{-e}, \mathbf{D}^*) \frac{I_G(\delta, \mathbf{D})}{I_{G'}(\delta, \mathbf{D})} \quad (8)$$

where, in general

$$H(d, e, \mathbf{K}^{-e}, \mathbf{S}) = \frac{I(d, S_{jj})}{J(d, S_{ee}, A_{11})} \left( \frac{|K_{V \setminus j}^0|}{|K_{V \setminus e}^1|} \right)^{(d-2)/2} \exp \left( -\frac{1}{2} \langle \mathbf{S}, K^0 - K^1 \rangle \right)$$

where  $I(b, c) = c^{-b/2} 2^{b/2} \Gamma(b/2)$ ,

$$J(h, \mathbf{B}, b) = \left( \frac{2\pi}{B_{22}} \right)^{1/2} b^{\frac{(h-1)}{2}} I(h, B_{22}) \exp \left( -\frac{b}{2} \left[ B_{11} - \frac{B_{12}^2}{B_{22}} \right] \right),$$

such that  $\mathbf{A} = \mathbf{K}_{ee} - \mathbf{K}_{e, V \setminus e} \mathbf{K}_{V \setminus e}^{-1} \mathbf{K}_{e, V \setminus e}$ . The matrix  $\mathbf{K}^0$  is equal to  $\mathbf{K}$  except that  $K_{jj}^0 = \mathbf{K}_{j, V \setminus j} \mathbf{K}_{V \setminus j}^{-1} \mathbf{K}_{j, V \setminus j}$  and  $K_{ij}^0 = K_{ji}^0 = 0$ . Finally, the matrix  $\mathbf{K}^1$  is equal to  $\mathbf{K}$  except that  $\mathbf{K}_e^1 = \mathbf{K}_{e, V \setminus e} \mathbf{K}_{V \setminus e}^{-1} \mathbf{K}_{e, V \setminus e}$ .

By using the CBF in (8), Wang and Li (2012) propose model moves that do not rely on RJ methods, and after assessing which graph to move to, the parameter  $K_{jj}$ , as well as  $K_{ij}$  if  $e$  is in the accepted graph, are resampled according to their conditional distributions given  $\mathbf{K}^{-e}$ . This method is appealing, as it offers an automatic manner of moving between graphs and does not rely on the tuning parameters used in the RJ methods of Dobra and Lenkoski (2011) and Dobra et al. (2011).

While the result has significant theoretical appeal we show that computation of the factor  $H(\delta + n, e, \mathbf{K}^{-e}, \mathbf{D}^*)$  is extremely costly, even in low dimensions. This is due to the formation of the matrices  $\mathbf{K}^0$  and  $\mathbf{K}^1$ , which require the solution of systems involving large matrices, in particular,  $\mathbf{K}_{V \setminus j}$  and  $\mathbf{K}_{V \setminus e}$ .

Suppose now that  $G$  and  $G'$  differ only by the edge  $f = (p-1, p)$  again with  $G \subset G'$ . We consider the CBF

$$\frac{pr(\mathcal{D}, \Phi^{-f}|G')}{pr(\mathcal{D}, \Phi^{-f}|G)},$$

where  $\Phi^{-f} = \Phi \setminus \{\Phi_{p-1, p}, \Phi_{pp}\}$ . In Appendix A we show that

$$\frac{pr(\mathcal{D}|\Phi^{-f}, G')}{pr(\mathcal{D}|\Phi^{-f}, G)} = N(\Phi^{-f}, \mathbf{D}^*) \frac{I_G(\delta, D)}{I_{G'}(\delta, D)}, \quad (9)$$

with, in general

$$N(\Phi^{-f}, \mathbf{S}) = \Phi_{p-1,p-1} \left( \frac{2\pi}{S_{pp}} \right)^{1/2} \exp \left( \frac{1}{2} S_{pp} (\phi_0 - \mu)^2 \right)$$

where  $\mu = \Phi_{p-1,p-1} S_{p-1,p} / S_{pp}$ , and  $\phi_0 = -\Phi_{p-1,p-1}^{-1} \sum_{l=1}^{p-2} \Phi_{lp-1} \Phi_{lp}$ .

This result originally appeared in an early version of Wang and Li (2012). In order to update a general edge  $e$ , we propose determining a permutation  $\vartheta$  of  $V$  such that the nodes of  $V \setminus e$  are reordered to reduce the fill-in of the graph  $G_{V \setminus e}$  and finally, the edge  $e$  is placed in the  $(p-1, p)$  position. Equation (9) is then calculated after permuting,  $\mathbf{K}$  and  $\mathbf{D}^*$  according to  $\vartheta$ .

The benefit of this method is the reduced computational overhead required to compute (9). The method requires relabeling the matrices  $\mathbf{K}$  and  $\mathbf{D}^*$  and determining the Cholesky decomposition of the permuted version of  $\mathbf{K}$ . Using node reordering software to minimize fill-in of  $G_{V \setminus e}$  proves useful in the developments below.

## 2.4 Avoiding Normalizing Constant Calculation

Both (8) and (9) require determination of the prior normalizing constants  $I_G$  and  $I_{G'}$ . While the MC method of Atay-Kayis and Massam (2005) enables these factors to be approximated, the routine can be subject to numerical instability (Lenkoski and Dobra, 2011; Wang and Li, 2012) and involves significant computational effort.

Wang and Li (2012) propose a method for avoiding the use of the MC approximation for prior normalizing constants. Their method employs the double Metropolis Hastings algorithm of Liang (2010), which is an extension of the exchange algorithm developed by Murray et al. (2006).

We briefly review the implementation of the double MH algorithm in Wang and Li (2012). Suppose that  $(K, G)$  is the current state of the MCMC chain and we propose to move to  $G'$  by adding the edge  $e$  to  $G$ . The double MH algorithm then forms a copy  $\tilde{\mathbf{K}}$  of  $\mathbf{K}$ , resamples  $\tilde{K}_{ij}$  and  $\tilde{K}_{jj}$  according to  $G'$ . It then updates  $\tilde{\mathbf{K}}$  via block Gibbs according to  $G'$ . Equation (8) is then replaced with

$$\frac{H(\delta + n, e, \mathbf{K}^{-e}, \mathbf{D}^*)}{H(\delta, e, \tilde{\mathbf{K}}^{-e}, \mathbf{D})} \quad (10)$$

We see that the expression (10) has replaced the prior normalizing constants with an evaluation of  $H$  in the prior, evaluated at  $\tilde{\mathbf{K}}$  (see Murray et al., 2006; Liang, 2010, for theoretical



justifications of this procedure). This is clearly beneficial, as it avoids the need for involved MC approximation. Unfortunately, the procedure as implemented in Wang and Li (2012) requires a full run of the block Gibbs sampler, as well as determination of the matrices  $\mathbf{K}^0$  and  $\mathbf{K}^1$  and therefore contains many large matrix operations.

We propose an alternative implementation of the double MH algorithm. Again suppose that  $(\mathbf{K}, G)$  is the current state and we propose to move to  $G'$  by adding the edge  $f = (p - 1, p)$ . We first determine  $\Phi$  from  $\mathbf{K}$ . We then update  $\Phi$  to  $\tilde{\Phi}$  using the RWMH algorithm in Section 2.2 relative to  $G'$ . Equation (9) is then replaced with

$$\frac{N(\Phi^{-f}, \mathbf{D}^*)}{N(\tilde{\Phi}^{-f}, \mathbf{D})} \quad (11)$$

We can immediately see that the expression (11) is considerably simpler than (10); it requires no additional matrix inversions nor the evaluation of any trace inner products. Furthermore, the generation of the auxiliary variables through the RWMH is considerably less demanding computationally than the use of the Block Gibbs sampler, especially when  $\mathbf{D} = \mathbb{I}_p$ , a common setting in practice.

## 2.5 Algorithms for Full Posterior Determination

In this section we outline the two algorithms we will consider for full posterior determination. Both algorithms create a sequence  $\{(\mathbf{K}^{[1]}, G^{[1]}), \dots, (\mathbf{K}^{[S]}, G^{[S]})\}$  where  $\mathbf{K}^{[s]} \in \mathbf{P}_{\mathbf{G}^{[s]}}$ . Given the current state  $(\mathbf{K}^{[s]}, G^{[s]})$  the WL algorithm proceeds as follows

0. Set  $\mathbf{K} = \mathbf{K}^{[s]}$  and  $G = G^{[s]}$

1. For each edge  $e$ , do:

a. if  $e \notin G$  attempt to update  $G$  to  $G' = G \cup e$  with probability

$$\frac{q(G'|\mathbf{K}^{-e}, \mathcal{D})}{q(G|\mathbf{K}^{-e}, \mathcal{D})} = \frac{pr(G')H(\delta + n, e, \mathbf{K}^{-e}, \mathbf{D}^*)}{pr(G)}$$

if  $e \in G$  the ratio is flipped. If  $G$  is not to be updated, skip to step c.

b. If attempting to update  $G$  to  $G'$ , sample  $\tilde{\mathbf{K}}$  as discussed in Section 2.4 and calculate

$$\alpha = \min\{1, H^{-1}(\delta, e, \tilde{\mathbf{K}}, \mathbf{D})\}$$

if  $e \in G'$ , otherwise calculate

$$\alpha = \min\{1, H(\delta, e, \tilde{\mathbf{K}}, \mathbf{D})\}$$

and with probability  $\alpha$  set  $G = G'$ , otherwise leave it unchanged.

c. Resample  $K_{ij}, K_{jj}$  according to  $G$ .

After attempting to update each edge, set  $G^{[s+1]} = G$ .

2. Resample  $\mathbf{K}^{[s+1]}$  using the block Gibbs sampler relative to  $G^{[s+1]}$  and the current state of  $\mathbf{K}$ .

We see that in one iteration of the WL algorithm, each edge is potentially updated in the graph. Our new algorithm (which we call CL) also follows this idea, and proceeds as follows

0. Set  $\mathbf{K} = \mathbf{K}^{[s]}$  and  $G = G^{[s]}$

1. For each edge  $e$ , do:

a. Determine a permutation  $\vartheta$  of  $V_p$  as discussed in Section 2.3, which places the edge  $e$  in the  $(p-1, p) = f$  position, and likewise permute  $\mathbf{K}$ ,  $G$ ,  $\mathbf{D}$  and  $\mathbf{D}^*$ . Let  $G^\vartheta$  denote the permuted version of  $G$  and  $\Phi$  be the Cholesky decomposition of the permuted version of  $\mathbf{K}$ . If  $f \notin G^\vartheta$  attempt to update  $G^\vartheta$  to  $G' = G^\vartheta \cup f$  with probability

$$\frac{q(G'|\Phi^{-f}, \mathcal{D})}{q(G^\vartheta|\Phi^{-f}, \mathcal{D})} = \frac{pr(G')N(\Phi^{-f}, D^*)}{pr(G^\vartheta)}$$

if  $f \in G^\vartheta$  the ratio is flipped. If  $G^\vartheta$  is not to be updated then, skip to step c.

b. If attempting to update  $G^\vartheta$  to  $G'$ , sample  $\tilde{\Phi}$  as discussed in Section 2.4 and calculate

$$\alpha = \min\{1, N^{-1}(\tilde{\Phi}^{-f}, \mathbf{D})\}$$

if  $f \in G'$ , otherwise calculate

$$\alpha = \min\{1, N(\tilde{\Phi}^{-f}, \mathbf{D})\}$$

and with probability  $\alpha$  set  $G^\vartheta = G'$ , otherwise leave it unchanged.

c. Resample  $\Phi_{p-1,p}, \Phi_{pp}$  according to  $G^\vartheta$ . Then reform  $\mathbf{K}$  and  $G$  by unpermuting the system.

After attempting to update each edge, set  $G^{[s+1]} = G$ .

2. Resample  $\mathbf{K}^{[s+1]}$  using the block Gibbs sampler relative to  $G^{[s+1]}$  and the current state of  $\mathbf{K}$ .

As we can see, there is somewhat more bookkeeping involved in the implementation of the CL algorithm, as the system is constantly being permuted. However, the reduction in computation time by using the RWMH algorithm and requiring only the calculation of the factors  $N(\Phi^{-f}, \mathbf{D}^*)$  and  $N(\tilde{\Phi}^{-f}, \mathbf{D})$  is dramatic, as we show below.

### 3 Simulation Study

In this section we conduct a simulation study that compares the method we have developed to the WL algorithm. Our example comes directly from Wang and Li (2012). We consider a situation in which  $p = 6$  and let  $\mathbf{U} = \mathbf{Y}\mathbf{Y}' = n\mathbf{A}^{-1}$  where  $n = 18$  and  $A_{ii} = 1$  for  $i = 1, \dots, 6$ ;  $A_{i,i+1} = A_{i+1,i} = .5$  for  $i = 1, \dots, 5$  and  $A_{16} = A_{61} = .4$ . We finally assume the prior  $\mathbf{K} \sim \mathcal{W}_G(3, \mathbb{I}_6)$ . Using exhaustive MC approximation of the entire graph space, Wang and Li (2012) show that the posterior probability of each edge is

$$(p_{ij}|A) = \begin{pmatrix} 1 & 0.969 & 0.106 & 0.085 & 0.113 & 0.85 \\ 0.969 & 1 & 0.98 & 0.098 & 0.081 & 0.115 \\ 0.106 & 0.98 & 1 & 0.982 & 0.0098 & 0.086 \\ 0.085 & 0.098 & 0.982 & 1 & 0.98 & 0.106 \\ 0.113 & 0.081 & 0.098 & 0.98 & 1 & 0.97 \\ 0.85 & 0.115 & 0.086 & 0.106 & 0.97 & 1 \end{pmatrix}$$

We use this example and compare the CL algorithm to the WL algorithm. Following Wang and Li (2012) we run both the WL and CL algorithms as described in Section 2.5 for 60,000 iterations and discard the first 10,000 iterations as burn-in. Both algorithms were implemented in R, though C++ was used for block-Gibbs updates. We note that if a pure R implementation had been used, the time differences between WL and CL would be even more dramatic.

We record the total computing time and looked at the mean squared errors of the posterior inclusion probabilities from the two runs compared with the true values given above. We

Table 1: Comparison of CL and WL algorithms for the six dimensional example.

	Time (sec)		MSE	
	Mean	SD	Mean	SD
CL	182.5	(4.1)	0.0088	(6e-04)
WL	818.4	(19.2)	0.0349	(0.0025)

repeated the entire process 100 times, each time starting both WL and CL from the same random starting point. Table 1 shows the average computing time in seconds (on a 2.8 GHz desktop computer with 4GB of RAM running Linux), average MSE and standard deviations across the 100 runs. The first column shows the expected result: even in six dimensions the WL algorithm takes more than 4 times as long to perform the same number of iterations as the CL algorithm. This shows the improved efficiency of the proposed method.

We found the results in the third column surprising, but do not draw broad conclusions from it. It appears that in this example, using 60,000 iterations, the CL algorithm approaches the true posterior edge expectation more quickly than the WL algorithm. Since both algorithms are correct theoretically, we choose not to emphasize this result. Furthermore, we have determined that by doubling the number of iterations, both approaches yield essentially the exact posterior distribution, though again the WL algorithm takes more than 4 times as long to run.

This example was chosen as it appears in Wang and Li (2012) and has an exact answer. The fact that the CL algorithm is considerably faster than the WL approach even in 6-dimensions indicates the broader appeal for searching truly high dimensional spaces.

## 4 A Multivariate Graphical Stochastic Volatility Model

Modeling the joint distribution of returns for a large number of assets is an important component of portfolio allocation and risk management. Carvalho and West (2007) and Rodriguez et al. (2011) both show that the use of GGMs can substantially improve modeling of joint asset returns. However, heterogeneity in asset returns was, at best, tangentially addressed. The study of Carvalho and West (2007) considered a fixed (decomposable) graph throughout the entire period and assumed that asset returns were identically distributed. Rodriguez et al.

(2011) allowed for mixing over the class of decomposable graphs and also introduced some heteroskedasticity by considering an infinite-dimensional hidden Markov model (iHMM). However, inside groups of observations in the iHMM, variances were assumed constant.

Despite these constraints in model assumptions, both studies showed substantial improvements by incorporating GGMs into the precision matrix associated with joint asset returns. We consider a situation in which the notion of homoskedastic, normally distributed asset returns is simply untenable; namely, the period surrounding the financial turbulence associated with the collapse of Lehman Brothers. We show that by utilizing the developments in the previous sections, we are able to specify a parsimonious stochastic volatility model for multivariate asset returns that quickly adapts to changes in market volatility. This model shows the flexibility of the new approach in embedding the GGM in larger hierarchical Bayesian frameworks.

## 4.1 The stochastic volatility model

Let  $\mathbf{Y}_t$  be the log-returns of  $p$  correlated assets. We specify the following hierarchical model for these returns

$$\begin{aligned} \mathbf{Y}_t | \mathbf{K}, X_t &\sim \mathcal{N}_p(\mathbf{0}, \exp(X_t) \mathbf{K}^{-1}) \\ X_t | \phi, X_{t-1}, \tau &\sim \mathcal{N}(\phi X_{t-1}, \tau^{-1}) \\ \phi &\sim \mathcal{N}(0, \tau_0) \\ \tau &\sim \Gamma(a, b) \\ K | G &\sim \mathcal{W}_G(\delta, \mathbf{D}) \\ G &\sim pr(G). \end{aligned} \tag{12}$$

In the likelihood (12) we see that asset returns are assumed to be mean-zero. The  $X_t$  terms then dictate an overall level of market volatility, while a constant precision parameter  $\mathbf{K}$  dictates the degree to which asset returns are correlated. While this model is parsimonious, it serves as a useful first departure from previous studies as it explicitly incorporates notions of stochastic volatility. For purposes of identification, we set  $X_0 = 0$ . In the conclusions section we discuss further possible generalizations to this framework.

After collecting a time-series of returns  $\mathbf{Y}^{(1:T)}$ , we then aim to determine the posterior distribution

$$pr(K, G, \tau, \phi, \mathbf{X} | \mathbf{Y}^{(1:T)})$$

where  $\mathbf{X} = (X_1, \dots, X_T)$ . Furthermore, we may be interested in the posterior predictive distribution  $pr(\mathbf{Y}^{(T+1)}|\mathbf{Y}^{(1:T)})$ . The parameters  $\mathbf{X}, \phi, \tau$  are updated with standard block MH or Gibbs steps (see Rue and Held, 2005) and the posterior predictive distribution is easily formed from these parameters. However, we note in particular that

$$\mathbf{K}|G, \tau, \phi, \mathbf{X}, \mathbf{Y}^{(1:T)} \sim \mathcal{W}_G \left( \delta + T, \mathbf{D} + \sum_{t=1}^T \frac{\mathbf{Y}^{(t)} \mathbf{Y}^{(t)'}}{\exp(X_t)} \right). \quad (13)$$

From (13) we see why the developments in Section 2 prove useful. We may update  $\mathbf{K}$  and  $G$  jointly using the CL algorithm discussed in Section 2.5 simply by setting  $\mathbf{D}^* = \mathbf{D} + \sum_{t=1}^T (\mathbf{Y}^{(t)} \mathbf{Y}^{(t)'}) / \exp(X_t)$ . This allows us to easily embed a sparse precision matrix  $\mathbf{K}$  and mix over the class of GGMs in any hierarchical Bayesian model that involves a standard Wishart distribution.

## 4.2 The data

To apply our model and algorithm we randomly chose 20 stocks from the S&P 500. These stocks were: Aetna Inc. (AET), CA Inc. (CA), Campbell Soup (CPB), CVS Caremark Corp. (CVS), Family Dollar Stores (FDO), Honeywell Int'l Inc. (HON), Hudson City Bancorp (HCBK), JDS Uniphase Corp. (JDSU), Johnson Controls (JCI), Morgan Stanley (MS), PPG Industries (PPG), Principal Financial Group (PFG), Sara Lee Corp. (SLE), Sempra Energy (SRE), Southern Co. (SO), Supervalu Inc. (SVU), Thermo Fisher Scientific (TMO), Wal-Mart Stores (WMT), Walt Disney Co. (DIS), Wellpoint Inc. (WLP).

We chose a time period where markets experience both high and low volatility to evaluate the flexibility of our model. We chose the time period from October 31, 2001 to May 21, 2008 as our training period to fit our model and make predictions for the time period from May 22, 2008 to October 23, 2009. The time periods consist of 1650 and 360 trading days respectively. Figure 1 shows the mean of the squared returns for the these 20 securities over the entire dataset. The extreme volatility present in the markets after the collapse of Lehman brothers in September 2008 is readily evident, showing that a homoskedasticity assumption is untenable for these data.

## 4.3 Predictive Performance Results

We assess the relative performance of the stochastic volatility model we develop in Section 4.1 versus a method that embeds GGMs, but does not have a stochastic volatility component.

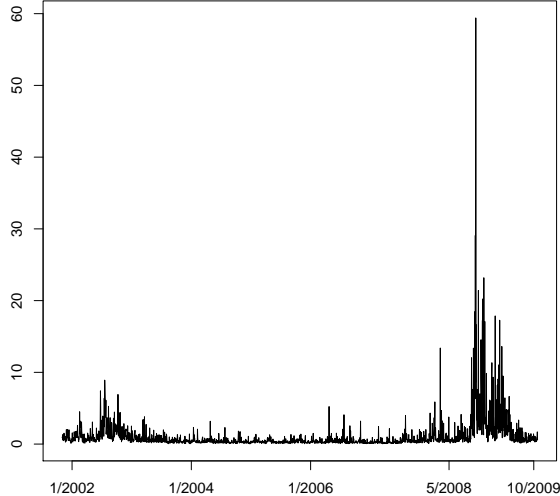


Figure 1: Mean of the squared returns taken over all 20 stocks during the entire time period from October 31, 2001 to October 23, 2009.

For each day  $t + 1$  in the forecast period we run our stochastic volatility model from the beginning of the training period until the previous day  $t$  to obtain estimates for the model parameters. We run the algorithm for 60,000 iterations, discarding the first 10,000 as burn-in (keep in mind that in one iteration of the algorithm all edges are evaluated). Using the posterior sample, we obtain the posterior predictive distribution of  $\mathbf{Y}^{(t+1)}$ .

Figure 2 shows the mean of the posterior predictive distribution of the volatility component  $X_{t+1}$  using the returns up to time  $t$ , which drives the predictive distribution of  $\mathbf{Y}^{(t+1)}$  for each day in the forecast period. Comparing Figures 1 and 2 we see that our model reflects the time-dependent volatility well. At the beginning when the market is quiet,  $X_{t+1}$  takes lower values mostly between 0 and 1. After the shock of the financial crisis the volatility in the market goes up extremely, which is reflected by significantly higher values of  $X_{t+1}$ . Months later, towards the end of the forecast period, the market has cooled down and the terms  $X_{t+1}$  reflect this.

The two methods we compare both return full predictive distributions. By construction, these predictive distributions have the same mean and median since returns are assumed mean-zero. Judging their performance therefore requires assessing the entire predictive distribution. We assess their performance using the energy score introduced by Gneiting and Raftery (2007).

The energy score is a proper scoring rule, which is a multivariate generalization of the

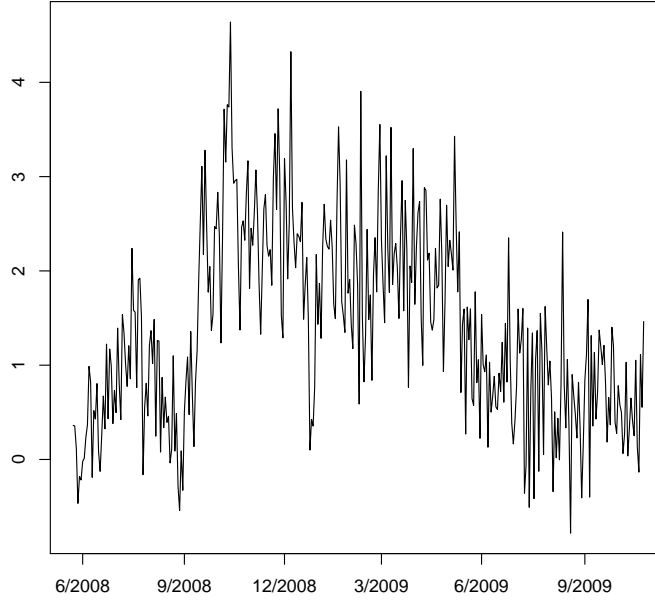


Figure 2: Means of posterior predictive distribution for the volatility component  $X_{t+1}$  from May 22, 2008 to October 23, 2009.

continuous ranked probability score. It is defined as

$$ES(F, \mathbf{x}) = \mathbf{E}_F \|\mathbf{X} - \mathbf{x}\|^2 - \frac{1}{2} \mathbf{E}_F \|\mathbf{X} - \mathbf{X}'\|^2 \quad (14)$$

where  $F$  is our predictive distribution for a vector-valued quantity,  $\mathbf{X}$  and  $\mathbf{X}'$  are independent random variables with distribution  $F$  and  $\mathbf{x}$  is the realization.

For each day in the forecast period, we compute the energy score for the predictive distribution returned by the two methods considered. Figure 3 shows the difference between the stochastic volatility model developed in Section 4.1 and the model that incorporates GGM uncertainty but holds volatility fixed. As we can see in Figure 3, between May and August, 2008, there is no clear difference between the two approaches. However, after the financial turbulence in September, 2008, the stochastic volatility model outperforms the fixed volatility model by a considerable margin. During almost every day in the turbulent period, the energy scores are lower under the stochastic volatility model. After the market turbulence subsides, the two models return to performing equally well again.

This short example shows the utility of the computational methodology developed in this paper. The model is simple, in many respects, but a non-trivial deviation from the standard iid sampling framework to which the GGM was initially relegated. By now being



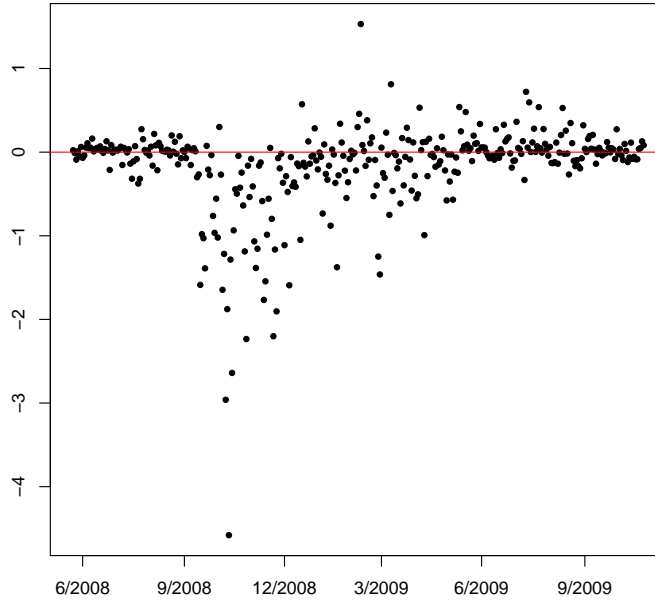


Figure 3: Difference of energy score from the predictive distribution of the model with stochastic volatility versus the model with fixed volatility. Values below zero indicate the stochastic volatility model outperformed the fixed volatility model.

able to embed the GGM in more complicated hierarchical frameworks, we are able to address difficult sampling schemes while simultaneously incorporating sparsity in the estimate of joint distributions.

## 5 Conclusions

We have synthesized a number of recent results related to the G-Wishart distribution. This has allowed for an algorithm that does not rely on RJ methods, obviates the need for expensive and numerically unstable MC approximation of prior normalizing constants and does so with minimal computational effort. The improvement in computation time is sufficient that at each stage of the algorithm, all edges may be evaluated for inclusion or exclusion in the graphical model. This algorithm allows the GGM to be embedded in more sophisticated hierarchical Bayesian models and opens the possibility of replacing standard Wishart distributions with G-Wishart variates, leveraging the improvement in predictive performance offered by sparse precision matrices.

The applied example shows the usefulness of this combination. We are able to sparsely model the interactions in financial assets while simultaneously addressing the issues of stochastic volatility prevalent in markets undergoing turbulence. The method is able to characterize the distribution of asset returns during periods of rapidly fluctuating volatility much better than standard iid frameworks.

The stochastic volatility model we develop remains parsimonious and several adjustments could be made. The first such development would be to replace the univariate term  $X_t$  with a multivariate factor that allows the variance of each asset to follow its own path, while potentially tying the evolution of these factors together with a separate GGM. Furthermore, employing some form of the iHMM framework of Rodriguez et al. (2011) could allow for the matrix  $\mathbf{K}$  to change throughout the period as well. Such developments will be considered in future work.

## References

- Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92:317–335.
- Carvalho, C. M. and West, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis*, 2:69–98.
- Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21:1272–1317.
- Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–175.
- Dickey, J. M. and Gunel, E. (1978). Bayes factors from mixed probabilities. *J. R. Statist. Soc. B*, 40:43–46.
- Dobra, A. and Lenkoski, A. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5:969–993.
- Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, *Forthcoming*.

- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102:359–378.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(711-732).
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20:388–400.
- Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*, 20:140–157.
- Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.*, 35:1278–323.
- Liang, F. (2010). A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computing and Simulation*, 80:1007–1022.
- Mitsakakis, N., Massam, H., and Escobar, M. D. (2011). A Metropolis-Hastings based method for sampling from the g-Wishart distribution in Gaussian graphical models. *Electronic Journal of Statistics*, 5:18–30.
- Murray, I., Ghahramani, Z., and MacKay, D. (2006). Mcmc for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*.
- Piccioni, M. (2000). Independence structure of natural conjugate densities to exponential families and the Gibbs Sampler. *Scand. J. Statist.*, 27:111–27.
- Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.*, 36:2818–2849.
- Rodriguez, A., Dobra, A., and Lenkoski, A. (2011). Sparse covariance estimation in heterogeneous samples. *Electronic Journal of Statistics*, 5:981–1014.
- Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.*, 29:391–411.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63:325–338.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall.

Wang, H. and Carvalho, C. M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics*, 4:1470–1475.

Wang, H. and Li, S. Z. (2012). Efficient Gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198.

## Appendix A: Determination of CBF for using $\Phi^{-f}$

Consider

$$\frac{pr(\mathcal{D}, \Phi^{-f}|G')}{pr(\mathcal{D}, \Phi^{-f}|G)}$$

we note that

$$pr(\mathcal{D}, \Phi^{-f}|G') = \int_{\Phi_{p-1,p}} \int_{\Phi_{pp}} pr(\mathcal{D}, \Phi|G') d\Phi_f$$

and

$$pr(\mathcal{D}, \Phi^{-f}|G) = \int_{\Phi_{pp}} pr(\mathcal{D}, \Phi|G) d\Phi_{pp}$$

up to common terms we thus have that

$$pr(\mathcal{D}, \Phi^{-f}|G') \propto \frac{\Phi_{p-1,p-1}}{I_{G'}(\delta, \mathbf{D})} \int_{\Phi_{p-1,p}} \exp\left(-\frac{1}{2}D_{p,p}^*(\Phi_{p-1,p} + \mu)^2\right) d\Phi_{p-1,p}$$

recognizing the integral as the kernel of a normal distribution, this yields

$$pr(\mathcal{D}, \Phi^{-f}|G') \propto \frac{\Phi_{p-1,p-1}}{I_{G'}(\delta, \mathbf{D})} \left(\frac{2\pi}{D_{pp}^*}\right)^{1/2}.$$

Further, again up to common terms

$$pr(\mathcal{D}, \Phi^{-f}|G) \propto \frac{1}{I_G(\delta, \mathbf{D})} \exp\left(-\frac{1}{2}D_{p,p}^*(\Phi_0 + \mu)^2\right)$$

and thus

$$\frac{pr(\mathcal{D}, \Phi^{-f}|G')}{pr(\mathcal{D}, \Phi^{-f}|G)} = N(\Phi^{-f}, D^*) \frac{I_G(\delta, \mathbf{D})}{I_{G'}(\delta, \mathbf{D})}$$